

From Feasibility to Ecosystems: How Generative AI at the Edge Has Evolved

Roberto Morabito
EURECOM
Biot, France
roberto.morabito@eurecom.fr

Danilo Pietro Pau
STMicroelectronics
Agrate, Italy
danilo.pau@st.com

Abstract—Generative AI at the edge has rapidly transitioned from a question of feasibility to the emergence of full-fledged ecosystems. Early efforts focused on demonstrating whether small language models could execute on constrained devices. In contrast, current systems emphasize distributed intelligence, agentic workflows, multimodal processing, and end-to-end software toolchains. This article distills insights gathered over four *Generative Edge AI forums* held between 2024 and 2025, organized by the *Edge AI Foundation*, involving 48 organizations across academia and industry. We analyzed recurring themes, technical shifts, and survey-based industry feedback, identifying five evolutionary waves that characterized the maturation of Generative Edge AI. Our findings suggest that the primary bottleneck has moved beyond model optimization toward system integration, reproducibility, and operational workflows. We conclude by outlining key implications for pervasive and edge computing research, highlighting open challenges in orchestrating heterogeneous devices, enabling multimodal intelligence, and building robust, developer-friendly edge AI ecosystems.

Index Terms—Generative Edge AI, Edge Computing, Small Language Models, Agentic AI Systems, Distributed Edge Intelligence, Multimodal AI, AI Toolchains, Pervasive Computing

I. INTRODUCTION

Generative artificial intelligence has traditionally been associated with centralized cloud infrastructures, where abundant computational and memory resources enable the execution of large-scale models. For a long time, this association shaped a widely held assumption: generative models, particularly language models, were fundamentally incompatible with resource-constrained edge devices.

Over the past few years, this assumption has progressively eroded. Advances in model compression, architectural efficiency, hardware acceleration, and inference optimization have enabled a growing class of sub-4B parameter language models, often referred to as Small Language Models (SLMs), to execute on embedded, mobile, and edge platforms [1].

Figure 1 illustrates the steady increase in the number and diversity of compact language models released over recent years. Rather than representing scaled-down variants of cloud models, many of these systems are explicitly designed for constrained environments, prioritizing memory footprint, latency, and energy efficiency. This trend reflects a broader shift in the language modeling community toward compact-by-design architectures optimized for on-device and near-device deployment.

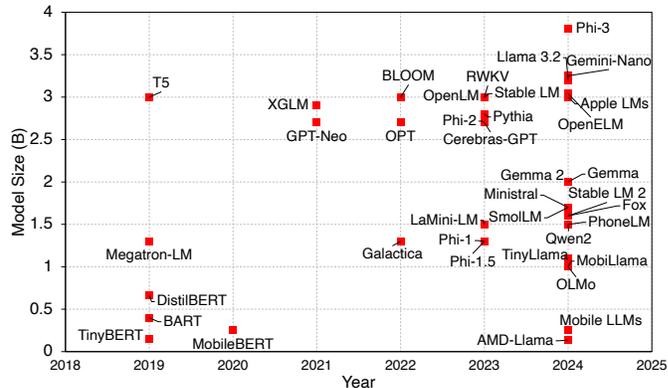


Fig. 1. Growth of sub-4B parameter language models over time. The figure illustrates the increasing availability and diversity of compact language models explicitly designed for efficiency, making them suitable for deployment on edge and mobile platforms.

At the same time, interest in Edge AI has evolved in parallel. While early Edge AI research focused primarily on perception tasks, such as vision, audio, and sensor-based inference, the emergence of capable SLMs has expanded the scope of intelligence that can be realistically placed at the edge. Notably, search trend data suggests that discussions around SLMs and Edge AI have increasingly followed a coupled trajectory [2].

As shown in Figure 2, periods of increased attention to SLMs coincide with rising interest in Edge AI. This temporal alignment raises an important question: *do advances in SLM capability act as a catalyst for renewed interest in Edge AI, or is Edge AI itself evolving into what can now be characterized as Generative Edge AI?*

Rather than viewing these interpretations as mutually exclusive, this article argues that the relationship is co-evolutionary. Improvements in SLM efficiency make generative workloads feasible on edge platforms, while the constraints and requirements of edge deployments, in turn, shape how generative models are designed, optimized, and integrated into systems. In this sense, Edge AI is not merely adopting generative capabilities; it is actively redefining how generative intelligence is engineered and operationalized.

Crucially, feasibility alone does not imply maturity. As generative models began to appear on edge devices, new chal-

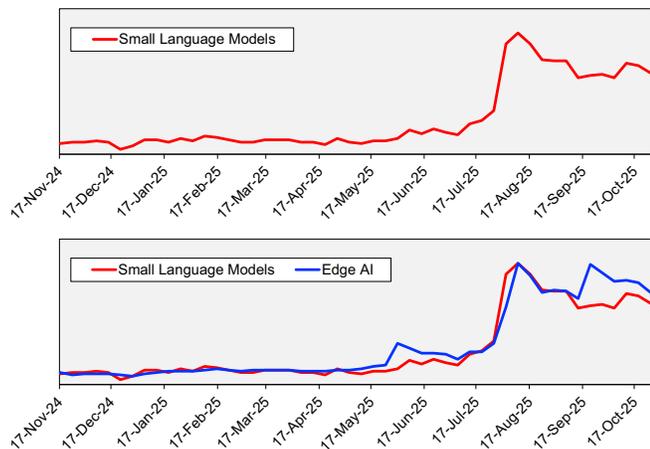


Fig. 2. Search trend comparison between “Small Language Models” and “Edge AI.” The paired rise suggests a close temporal relationship between advances in compact language model capabilities and increased attention to edge AI, indicating a co-evolution toward Generative Edge AI.

allenges emerged that extend beyond raw inference performance. Issues related to system integration, orchestration across heterogeneous devices, multimodal data fusion, and reproducible development workflows have become increasingly prominent. These challenges suggest a shift from model-centric optimization toward system-level design considerations.

This article examines how Generative Edge AI has evolved from early feasibility demonstrations to the emergence of broader ecosystems. Drawing on insights collected from four Edge AI Foundation Generative Edge AI forums held between 2024 and 2025 (bringing together 48 organizations across academia and industry), we analyze recurring themes, technical shifts, and industry perspectives. Our goal is to characterize this evolution, identify its driving forces, and highlight the implications for future pervasive and edge computing research.

Methodology and Community Perspective

Our analysis is informed by a longitudinal observation of four Generative Edge AI forums spanning an 18-month period. These events brought together 48 organizations, including universities, semiconductor vendors, startups, large technology companies, and applied research laboratories (Figure 3). Unlike single-shot workshops, the repeated nature of these forums enabled the identification of persistent themes, emerging priorities, and shifting technical emphases over time.

In addition to qualitative observations, we conducted a survey among Edge AI Foundation partners to capture expectations regarding adoption timelines, perceived barriers, and anticipated impact of Generative Edge AI solutions. While the survey is not statistically representative of the broader market, it provides indicative signals from organizations actively developing and deploying edge AI systems.

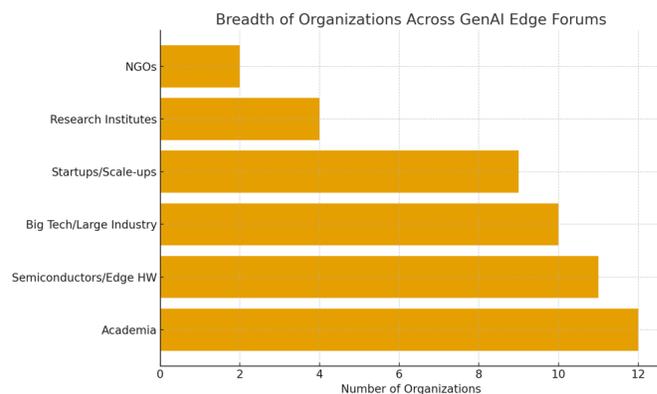


Fig. 3. Breadth of organizations participating across four Generative Edge AI forums (2024–2025). The distribution highlights strong engagement from academia, semiconductor and edge hardware vendors, large technology companies, startups, research institutes, and NGOs, reflecting the cross-disciplinary and ecosystem-driven nature of Generative Edge AI.

II. FIVE WAVES IN THE EVOLUTION OF GENERATIVE EDGE AI

Drawing from four Generative Edge AI forums held between 2024 and 2025, we identify five recurring and progressively dominant themes that characterize the evolution of Generative Edge AI. Rather than representing strictly sequential stages, these waves overlap in time and influence one another. Together, they capture a clear shift from feasibility-driven experimentation toward system-level maturity and ecosystem formation.

A. Wave 1: On-Device Feasibility — “Can We Run SLMs?”

The first wave was dominated by a fundamental feasibility question: can SLMs run on edge devices at all? Early discussions and demonstrations focused almost exclusively on porting, optimizing, and executing sub-billion and low-billion parameter models on constrained hardware platforms [3]. The primary technical challenges revolved around (i) aggressive model compression and quantization, (ii) reduction of memory footprint to fit within limited on-device resources, and (iii) achieving acceptable inference latency on CPUs and embedded accelerators.

Representative efforts during this phase broadly explored the execution of compact language models on constrained edge platforms, often through tightly scoped, device-specific implementations. These efforts typically emphasized practical feasibility over generality, focusing on single-device setups and controlled evaluation scenarios. Despite their limited scope, they played a critical role in challenging the prevailing assumption that generative models necessarily belonged in the cloud. By the end of this wave, on-device feasibility had shifted from an open question to a partially resolved problem, enabling the community to move beyond proof-of-concept demonstrations.

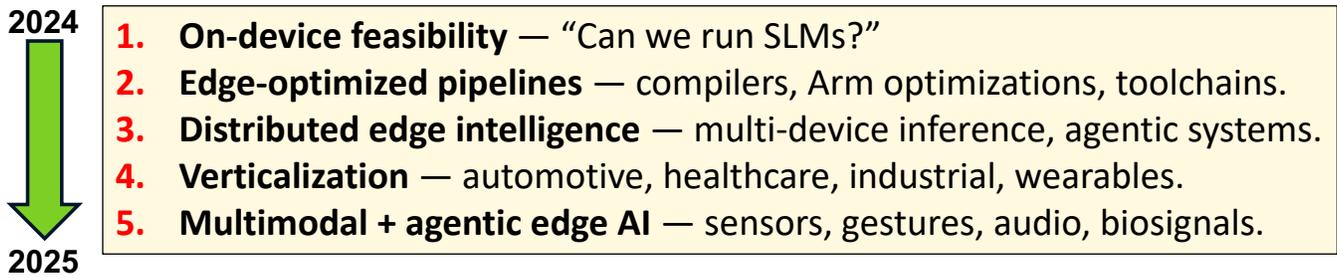


Fig. 4. Five evolutionary waves of Generative Edge AI observed across four Edge AI Foundation forums (2024–2025). The progression highlights a shift from on-device feasibility of SLMs toward distributed, multimodal, and workflow-centric edge AI ecosystems, where tooling, system integration, and domain specialization increasingly dominate research and development efforts.

B. Wave 2: Edge-Optimized Pipelines — From Models to Toolchains

As feasibility concerns began to recede, attention shifted toward how generative models are executed at the edge. The second wave emphasized inference pipelines, compiler support, and hardware-aware optimizations, marking an important transition from model-centric to pipeline-centric thinking [4]. Discussions increasingly focused on (i) end-to-end inference pipelines rather than isolated kernels, (ii) compiler optimizations tailored to Arm and embedded architectures, and (iii) runtime systems that efficiently expose hardware acceleration.

Across this phase, a growing body of work highlighted the importance of software infrastructure in determining generative inference performance at the edge. Rather than attributing efficiency gains solely to model architecture, these efforts increasingly emphasized the role of end-to-end pipelines, compiler support, and hardware-aware runtimes. This shift reflected an emerging consensus that performance, portability, and maintainability depend critically on the maturity of the surrounding software stack. As a result, toolchains began to be recognized as first-class components of Generative Edge AI systems rather than auxiliary implementation details.

C. Wave 3: Distributed Edge Intelligence — Multi-Device and Agentic Systems

The third wave marked a significant conceptual leap from optimizing single devices to coordinating intelligence across multiple edge nodes [5]. As individual devices became capable of running generative models, new opportunities emerged for collaborative and distributed inference [2], [6]. Key themes included (i) partitioning generative workloads across heterogeneous devices, (ii) coordinating multiple SLMs with complementary capabilities, and (iii) the emergence of early agentic frameworks designed specifically for edge environments.

During this phase, generative inference increasingly began to be reframed as a system-level orchestration problem rather than a standalone execution task. Research efforts explored how multiple compact models could cooperate across heterogeneous edge nodes, shifting attention toward coordination, communication, and runtime interaction among devices.

This transition reintroduced classical distributed systems challenges, including synchronization, communication overhead, and fault tolerance, into the context of generative AI, highlighting the need for new abstractions tailored to edge deployments [7].

D. Wave 4: Verticalization — From Generic Demos to Industry-Grade Applications

While early forums were dominated by generic demonstrations of “GenAI on the edge,” subsequent discussions showed a clear shift toward vertically specialized applications. This fourth wave reflected growing industry engagement and the realization that real-world deployments impose domain-specific constraints. Prominent verticals included (i) automotive systems [8], where safety, latency, and robustness are paramount, (ii) healthcare and biosensing applications [9], characterized by strict privacy, energy, and signal quality requirements, and (iii) industrial environments that integrate modern AI capabilities with legacy infrastructure [10].

Across this phase, a growing emphasis on vertically specialized applications became evident, reflecting the increasing involvement of industry stakeholders and real-world deployment scenarios. Rather than generic demonstrations, efforts in this wave focused on meeting domain-specific requirements and integrating generative capabilities into existing operational workflows. As a result, success was no longer measured primarily by benchmark performance, but by the ability to satisfy application constraints and achieve seamless integration within established systems. This verticalization marked a decisive step toward practical adoption.

E. Wave 5: Multimodal and Workflow-Centric Edge AI — From Vision to Engineering

The most recent wave reflects a shift from high-level research vision toward concrete engineering practices. Discussions increasingly emphasize multimodal intelligence [11], software architectures, and complete development workflows. Key characteristics include (i) integration of language models with vision, audio, and sensor data, (ii) cross-modal reasoning on constrained devices, and (iii) explicit attention to software architecture, deployment workflows, and reproducibility.

TABLE I
SUMMARY OF THE FIVE EVOLUTIONARY WAVES OF GENERATIVE EDGE AI, HIGHLIGHTING THEIR PRIMARY FOCUS, DOMINANT CHALLENGES, AND REPRESENTATIVE THEMES.

Wave	Period	Primary Focus	Key Challenges and Themes
1	2024	On-device feasibility	Running SLMs on constrained hardware; memory footprint, latency, and basic runtime support
2	2024–2025	Edge-optimized pipelines	Compiler support, hardware-aware optimizations, end-to-end inference pipelines
3	2025	Distributed intelligence	Multi-device inference, agentic systems, coordination across heterogeneous nodes
4	2025	Verticalized applications	Domain-specific constraints in automotive, healthcare, industrial, and wearable systems
5	2025+	Workflow-centric systems	Multimodality, software architectures, reproducibility, and complete toolchains

Recent work in this phase highlights a clear transition from high-level research vision toward deployable systems, particularly in the context of multimodal edge intelligence [12]–[14]. Rather than isolated demonstrations, these efforts emphasize the engineering practices required to integrate generative models with sensing, perception, and actuation pipelines under real-world constraints. As a result, tooling, software architecture, and end-to-end workflow design increasingly determine success. This wave underscores that the primary bottleneck for Generative Edge AI is no longer model capability, but the engineering effort required to operationalize generative intelligence at scale.

Summary

Taken together (Table I), these five waves illustrate a broader transformation: Generative Edge AI is no longer defined by the question of whether it can be done, but by how effectively it can be engineered, integrated, and sustained. The dominant challenges are shifting toward system design, orchestration, and ecosystem-level considerations, aligning closely with long-standing research themes in pervasive and edge computing.

III. INDUSTRY PERSPECTIVES: ADOPTION, BARRIERS, AND IMPACT

Beyond qualitative observations from forum discussions, we collected feedback from industry and academic partners of the Edge AI Foundation to better understand expectations surrounding the adoption of Generative Edge AI [15]. The survey aimed to capture indicative trends related to market readiness, perceived technical barriers, and anticipated impact across application domains. While the survey does not claim statistical representativeness, it reflects the perspectives of organizations actively engaged in developing, deploying, or enabling edge AI systems.

Table II summarizes the main signals emerging from the survey. A consistent outcome is that Generative Edge AI is largely perceived as a near-term opportunity rather than a long-term research vision. Many respondents anticipate initial

TABLE II
INDUSTRY SURVEY SNAPSHOT SUMMARIZING INDICATIVE EXPECTATIONS AND PERCEIVED CHALLENGES RELATED TO GENERATIVE EDGE AI ADOPTION.

Dimension	Indicative Outcome
Expected adoption timeline	Near-term
Main adoption barriers	Tooling, system integration
High-impact sectors	Healthcare, industrial
Model types	SLMs, cross-modal models

deployments within relatively short timeframes, particularly in controlled, domain-specific environments where edge intelligence can deliver immediate value. This perception aligns with the rapid progress observed in SLMs, edge-optimized toolchains, and multimodal pipelines discussed in the previous section.

At the same time, respondents consistently emphasized that the main barriers to adoption are no longer dominated by raw model performance. Instead, challenges related to tooling and system integration were cited most frequently. These include difficulties in assembling heterogeneous software stacks, integrating generative models with existing edge pipelines, and managing interactions across devices with diverse computational capabilities. Such concerns reinforce the observation that Generative Edge AI has entered a phase where system-level maturity, rather than algorithmic novelty, is the primary determinant of practical viability.

Another notable signal from the survey concerns expectations around model types. Rather than emphasizing large, general-purpose language models, respondents highlighted the relevance of SLMs and cross-modal systems tailored to specific tasks and deployment contexts. This preference reflects the realities of edge environments, where predictability, resource efficiency, and tight coupling with sensing and actuation pipelines are often more valuable than broad generality.

From an application perspective, healthcare and industrial domains emerged as particularly promising early adopters. These sectors share a strong demand for localized intelligence, low-latency decision making, and reduced dependence

TABLE III
 OUTLOOK FOR GENERATIVE EDGE AI: KEY DIMENSIONS SHAPING THE TRANSITION FROM MODEL-CENTRIC RESEARCH TO ECOSYSTEM-LEVEL DEPLOYMENT.

Outlook Dimension	Key Implications and Research Directions
Operationalization	Focus on deployment, maintenance, and reproducibility rather than new model architectures
Agentic workflows	Edge-tailored coordination protocols accounting for resource constraints and device heterogeneity
Interoperability	Lightweight, hardware-aware mechanisms for model and agent interaction across platforms
Multimodality	Integration of language, vision, audio, and sensor data, often via domain-specific foundation models
Edge-friendly adaptation	Resource-efficient retrieval-augmented generation and lightweight fine-tuning techniques
Tooling and workflows	Developer-friendly pipelines that abstract hardware diversity and enable reproducible deployments

on continuous cloud connectivity. At the same time, they impose stringent requirements related to safety, reliability, and privacy, which further amplify the importance of robust system integration and reproducible deployment practices.

Overall, the survey findings complement the evolutionary trends identified in Section 3. Together, they suggest that the central challenge facing Generative Edge AI is not the absence of capable models, but the lack of standardized, reproducible, and developer-friendly workflows that bridge research prototypes and production systems. Addressing this gap represents a key opportunity for the pervasive computing community and motivates a renewed focus on system architectures, tooling, and end-to-end engineering practices.

IV. FROM MODELS TO ECOSYSTEMS: IMPLICATIONS AND OUTLOOK FOR GENERATIVE EDGE AI

The evolution observed across the Generative Edge AI forums suggests that the central research challenges in this area are shifting decisively away from model-centric optimization toward system-level design and ecosystem development, a trend that has also been highlighted in recent community-wide perspectives on Generative Edge AI [16].

A first and fundamental implication is that the next phase of Generative Edge AI will be defined less by the introduction of new model architectures and more by the ability to operationalize what is already available. SLMs and compact multimodal models have reached a level of maturity that enables meaningful deployment at the edge. The challenge now lies in transforming these capabilities into reliable, maintainable, and reproducible systems that can function under real-world constraints.

One key dimension of this transition concerns agentic workflows spanning multiple devices. As discussed in Section 3, recent efforts increasingly frame generative inference as a cooperative process involving multiple models and nodes. While significant momentum exists around agentic paradigms and interaction protocols in cloud and desktop settings [17], these approaches are not directly transferable to edge environments. Edge-tailored agentic systems must explicitly account for intermittent connectivity, constrained memory and energy budgets, and the heterogeneity of AI acceleration hardware.

Designing lightweight, interoperable agentic protocols that enable coordination without imposing prohibitive overhead remains an open research challenge.

Closely related to agentic workflows is the issue of interoperability. Emerging standards and frameworks for agent-to-agent communication and model interaction demonstrate strong ecosystem momentum [17]. However, edge deployments introduce additional dimensions that are often abstracted away in cloud-centric designs. These include differences in instruction set architectures, availability of specialized accelerators, and variability in runtime support across devices. Future Generative Edge AI systems will require interoperability mechanisms that are not only semantically aligned, but also resource-aware and hardware-conscious.

Another critical implication concerns the growing role of multimodality. As Generative Edge AI moves beyond text-centric use cases, the integration of language models with vision, audio, and sensor data becomes increasingly central. This trend is particularly evident in vertical domains such as healthcare, industrial monitoring, and automotive systems. In these contexts, multimodality is not an optional enhancement but a fundamental requirement. The emergence of domain-specific foundation models at the edge further reinforces this direction, enabling tighter coupling between generative reasoning and domain knowledge while respecting deployment constraints.

In parallel, there is renewed interest in edge-friendly approaches to model adaptation, including retrieval-augmented generation and lightweight fine-tuning. Unlike cloud-based settings, edge deployments must carefully balance adaptation benefits against storage, latency, and energy costs. This creates opportunities for research into selective, context-aware retrieval mechanisms and incremental adaptation techniques that operate within tight resource envelopes. Importantly, such approaches must integrate seamlessly with broader system workflows rather than existing as isolated optimization steps.

Across all these dimensions, tooling and reproducibility emerge as decisive factors. Survey results and forum discussions consistently highlight the difficulty of assembling end-to-end Generative Edge AI systems from heterogeneous components. As a result, tooling that simplifies deployment, abstracts

hardware diversity, and supports reproducible workflows is likely to become a key competitive differentiator. In this sense, Generative Edge AI is entering a phase where the quality of the ecosystem surrounding models may matter more than the models themselves.

Table III summarizes the main outlook dimensions and associated research directions discussed in this section. Together, they highlight a clear shift toward ecosystem-level thinking, where progress depends on coordinated advances across models, systems, and tools.

These observations indicate that Generative Edge AI is transitioning from a technology-driven exploration phase to an ecosystem-building phase. Addressing the outlined challenges requires a holistic perspective that aligns models, systems, and tools with the realities of pervasive and edge computing. The Edge AI Foundation community, with its breadth of stakeholders and practical focus, is well positioned to act as a catalyst in shaping this next phase.

V. CONCLUSION

Generative Edge AI has undergone a rapid transformation over a short period of time. What initially began as an exploration of feasibility, i.e., focused on whether SLMs could run on constrained devices, has evolved into a broader effort to build complete, deployable ecosystems at the edge. Through the analysis of four Generative Edge AI forums and complementary industry feedback, this article has highlighted a clear shift from model-centric experimentation toward system-level integration, orchestration, and operational maturity. Our findings suggest that the primary challenges facing Generative Edge AI are no longer rooted in model capability, but in the engineering effort required to integrate generative intelligence into heterogeneous, resource-constrained environments. Issues such as agentic coordination across devices, multimodal processing, interoperability, and reproducible tooling now dominate both research and industrial agendas. Addressing these challenges requires perspectives and methodologies that are deeply aligned with the goals of pervasive and edge computing.

Looking forward, progress in Generative Edge AI will depend on the ability to align models, systems, and tools into coherent and developer-friendly ecosystems. In this context, the success of future solutions will be determined less by individual model advances and more by the robustness, usability, and interoperability of the platforms that support them.

REFERENCES

- [1] M. Abstreiter, S. Tarkoma, and R. Morabito, "Sometimes painful but certainly promising: Feasibility and trade-offs of language model inference at the edge," *arXiv preprint arXiv:2503.09114*, 2025.
- [2] R. Morabito and S. Jang, "Smaller, smarter, closer: The edge of collaborative generative ai," *IEEE Internet Computing*, 2025.
- [3] S. Laskaridis, K. Katevas, L. Minto, and H. Haddadi, "Melting point: Mobile evaluation of language transformers," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 890–907.
- [4] T. Sipola, J. Alatalo, T. Kokkonen, and M. Rantonen, "Artificial intelligence in the iot era: A review of edge ai hardware and software," in *2022 31st Conference of Open Innovations Association (FRUCT)*. IEEE, 2022, pp. 320–331.

- [5] D. Yu, Z. Xie, Y. Yuan, S. Chen, J. Qiao, Y. Wang, Y. Yu, Y. Zou, and X. Zhang, "Trustworthy decentralized collaborative learning for edge intelligence: A survey," *High-Confidence Computing*, vol. 3, no. 3, p. 100150, 2023.
- [6] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," *IEEE Communications Surveys & Tutorials*, 2025.
- [7] S. Nath, C. Peridis, E. Benjamin, X. Liu, S. Kolouri, P. Kinnell, Z. Li, C. Liu, S. Dora, and A. Soltoggio, "Collaborative learning in agentic systems: A collective ai is greater than the sum of its parts," *arXiv preprint arXiv:2506.05577*, 2025.
- [8] L. Bischoff and M. Stephan, "Generative artificial intelligence as driver for innovation in the automotive industry—a systematic analysis," in *New Players in Mobility: Technische und betriebswirtschaftliche Aspekte*. Springer, 2025, pp. 77–98.
- [9] R. Qureshi, M. Irfan, H. Ali, A. Khan, A. S. Nittala, S. Ali, A. Shah, T. M. Gondal, F. Sadak, Z. Shah *et al.*, "Artificial intelligence and biosensors in healthcare and its clinical relevance: A review," *IEEE access*, vol. 11, pp. 61 600–61 620, 2023.
- [10] S. Sai, R. Sai, and V. Chamola, "Generative ai for industry 5.0: analyzing the impact of chatgpt, dalle, and other models," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 3056–3066, 2024.
- [11] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang *et al.*, "Towards artificial general intelligence via a multimodal foundation model," *Nature Communications*, vol. 13, no. 1, p. 3094, 2022.
- [12] A. A. Khan, A. A. Laghari, M. Alsafyani, A. M. Baqasah, N. Kryvinska, A. Almadhor, R. Alroobaea, and M. Gregus Jr, "A cost-effective approach using generative ai and gamification to enhance biomedical treatment and real-time biosensor monitoring," *Scientific Reports*, vol. 15, no. 1, p. 17305, 2025.
- [13] A. Gibbons, E. King, I. Donohue, and A. Parnell, "Generative ai-based data augmentation for improved bioacoustic classification in noisy environments," *arXiv preprint arXiv:2412.01530*, 2024.
- [14] M. S. Sarowar, N. E. J. Farjana, M. A. I. Khan, M. A. Mutalib, S. Islam, and M. Islam, "Hand gesture recognition systems: A review of methods, datasets, and emerging trends," *International Journal of Computer Applications*, vol. 975, p. 8887.
- [15] R. Morabito, R. Adorante, H. Mousannif, and D. P. Pau, "Expanding the horizons of generative edge ai: Mission, vision, and insights from industries," *Authorea Preprints*, 2025.
- [16] V. J. Reddi, "Generative ai at the edge: Challenges and opportunities: The next phase in ai deployment," *Queue*, vol. 23, no. 2, pp. 79–137, 2025.
- [17] Y. Yang, H. Chai, Y. Song, S. Qi, M. Wen, N. Li, J. Liao, H. Hu, J. Lin, G. Chang *et al.*, "A survey of ai agent protocols," *arXiv preprint arXiv:2504.16736*, 2025.